DISTRIBUTIONS OF CONTINUOUS DATA

ST101 – DR. ARIC LABARR

RANDOM VARIABLES

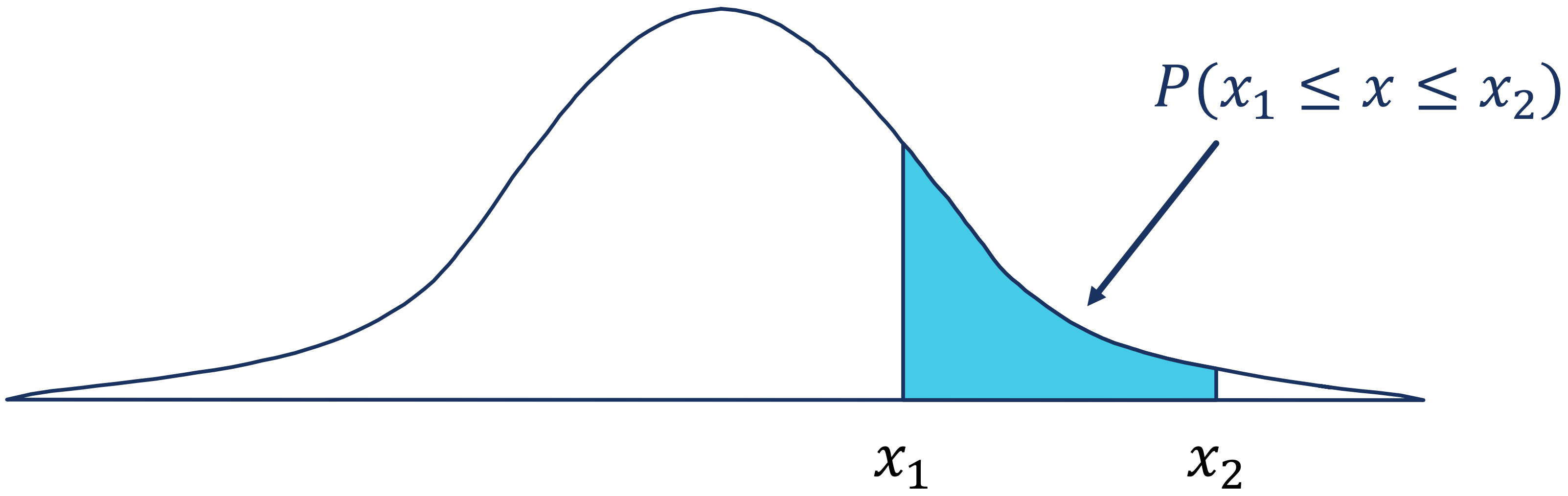
- A **random variable** is a numerical description of the outcome of an experiment.
- They can be either discrete or continuous.
- A discrete random variable may assume either a finite number of values or an infinite sequence of values.
- A **continuous random variable** may assume any numerical value in an interval or collection of intervals.

CONTINUOUS RANDOM VARIABLES

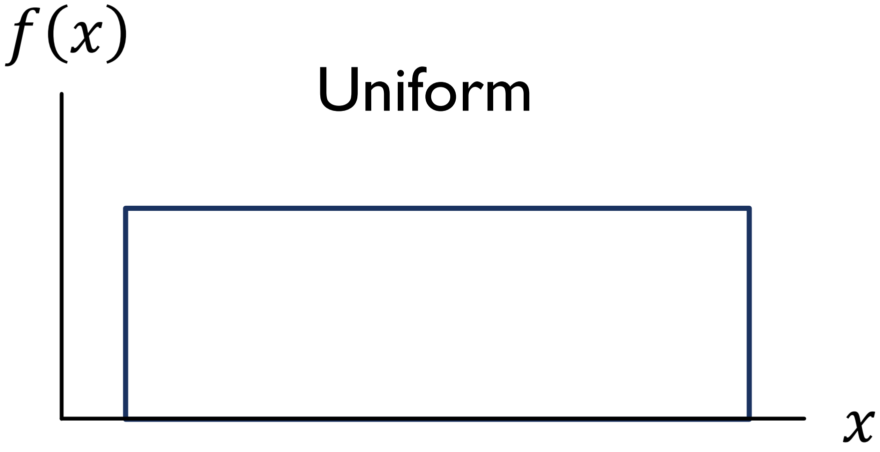
- A **continuous random variable** can assume any value in an interval on the real line or in a collection of intervals on the real line.
- It is **not** possible to talk about the probability of the random variable assuming a particular value.
- Instead, we talk about the probability of the random variable assuming a value inside of a given interval.

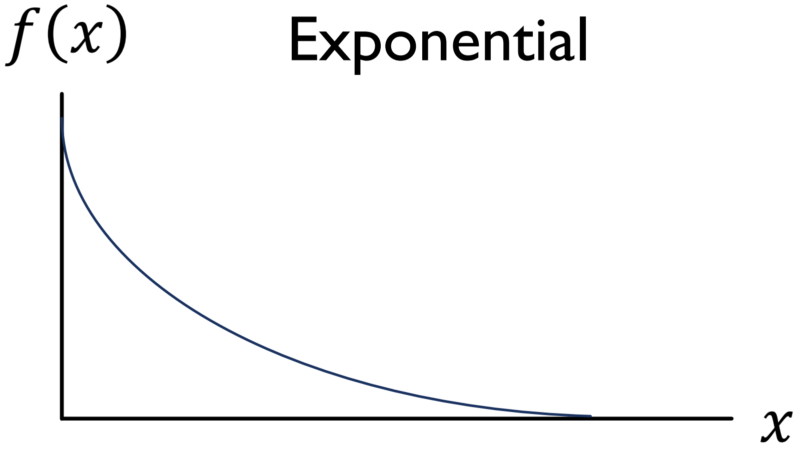
PROBABILITIES ON INTERVALS

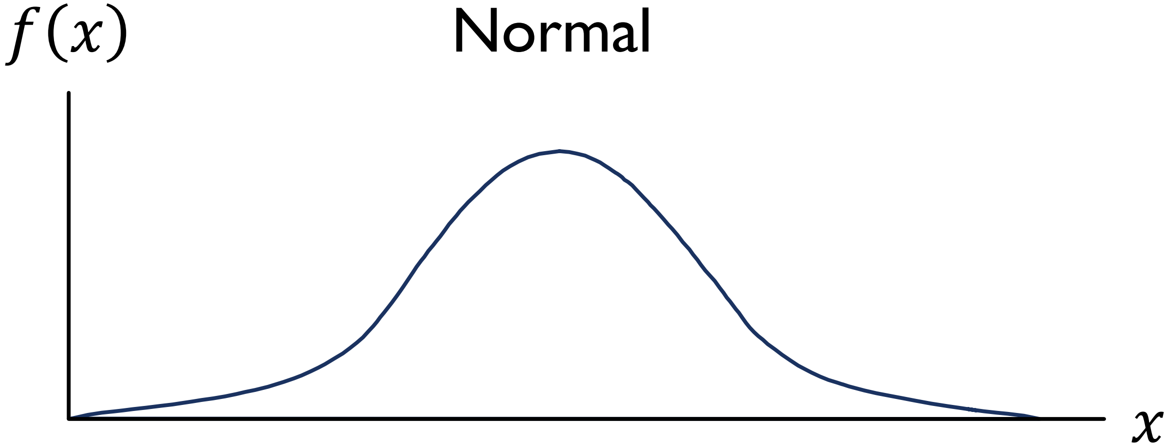
- Instead, we talk about the probability of the random variable assuming a value inside of a given interval.
- The probability of the random variable assuming a value inside of a given interval from x_1 to x_2 is the **area under the graph** of the **probability density function** between x_1 and x_2.



POPULAR CONTINUOUS DISTRIBUTIONS







SUMMARY

- A continuous random variable may assume any numerical value in an interval or collection of intervals.
- It is not possible to talk about the probability of the random variable assuming a particular value, but we instead talk about probabilities of intervals.

UNIFORM DISTRIBUTION

DISTRIBUTIONS OF CONTINUOUS DATA

UNIFORM PROBABILITY DISTRIBUTION

- A random variable follows a **uniform distribution** whenever the probability is proportional to the interval's length.
- In other words, every value has an equal probability of happening.
- The **probability density function** for the uniform distribution is:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0, & \text{elsewhere} \end{cases}$$

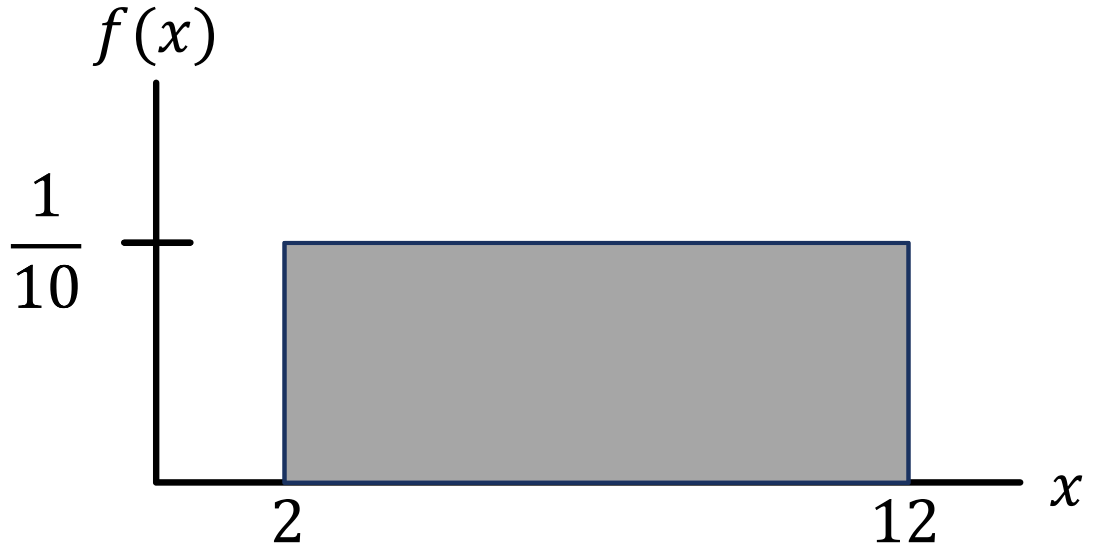
EXAMPLE OF UNIFORM DISTRIBUTION

- Assume that sales calls that go into a company are uniformly distributed by the years of experience of the sales staff so that everyone has the same chance of getting a call.
- The years of experience ranges from 2-12.

$$f(x) = \begin{cases} \dfrac{1}{12-2} = \dfrac{1}{10}, & 2 \le x \le 12 \\ 0, & \text{elsewhere} \end{cases}$$

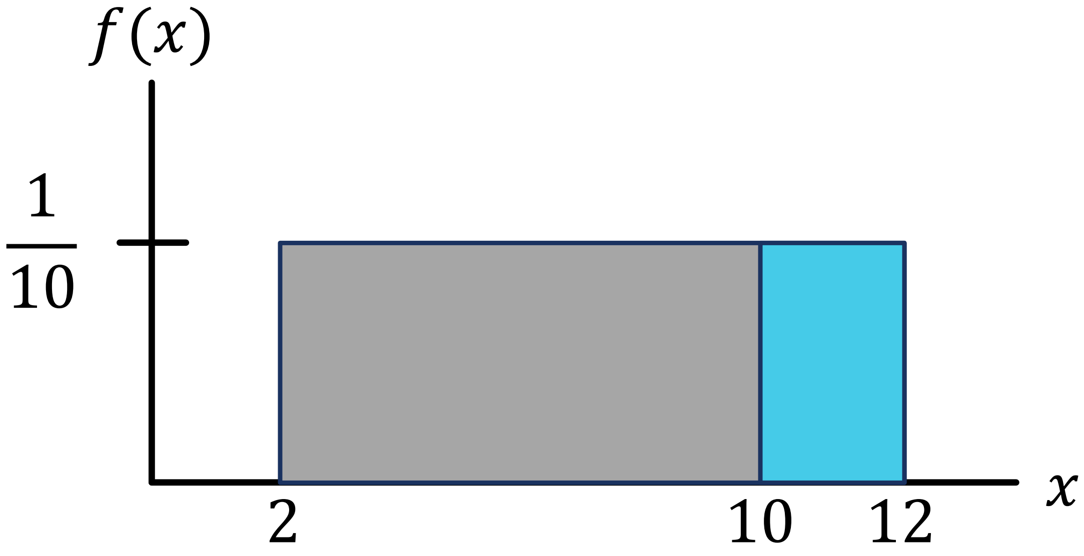
EXAMPLE OF UNIFORM DISTRIBUTION

- Assume that sales calls that go into a company are uniformly distributed by the years of experience of the sales staff so that everyone has the same chance of getting a call.



EXAMPLE OF UNIFORM DISTRIBUTION

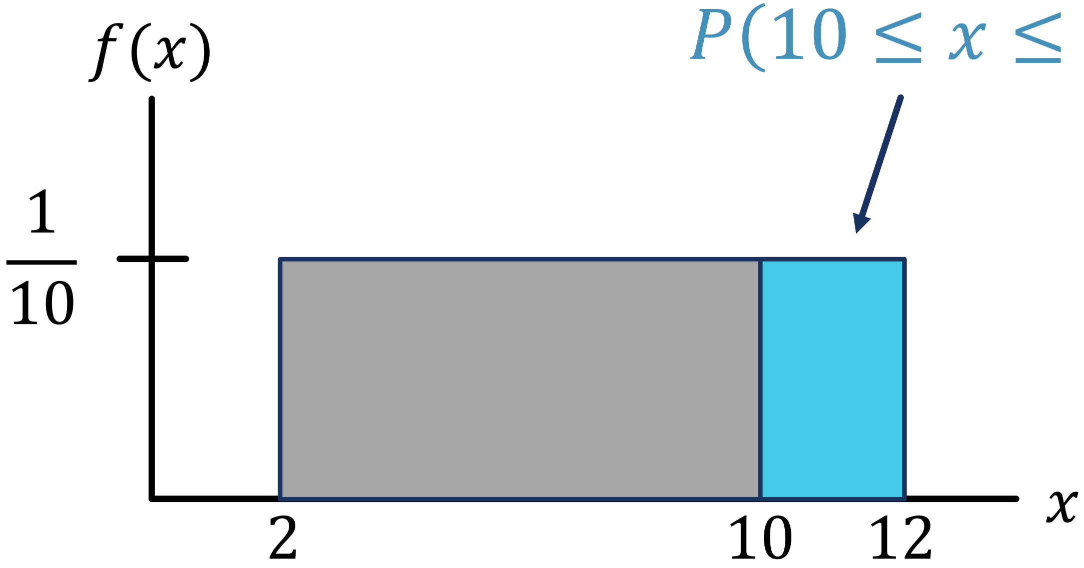
- Assume that sales calls that go into a company are uniformly distributed by the years of experience of the sales staff so that everyone has the same chance of getting a call.
- What is the probability a call is answered by an employee with 10 to 12 years of experience?



EXAMPLE OF UNIFORM DISTRIBUTION

- What is the probability a call is answered by an employee with 10 to 12 years of experience?
- Area under the curve between 10 and 12.

$$P(10 \leq x \leq 12) = 2 \times \frac{1}{10} = 0.2$$



MEASURES ON UNIFORM DISTRIBUTION

- Expected Value:

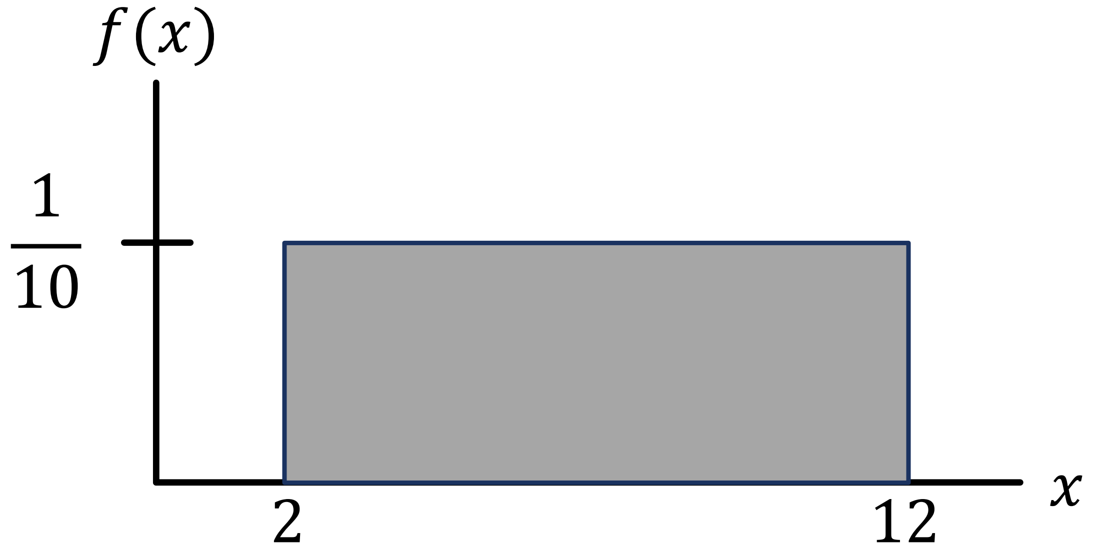
$$E(x) = \frac{a+b}{2}$$

- Variance:

$$Var(x) = \frac{(b-a)^2}{12}$$

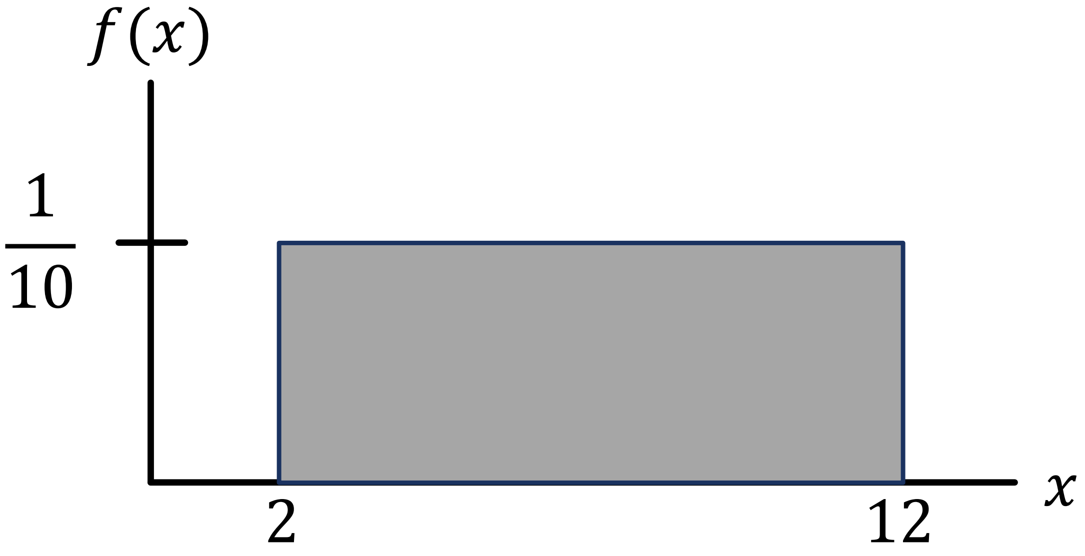
EXAMPLE OF UNIFORM DISTRIBUTION

- Assume that sales calls that go into a company are uniformly distributed by the years of experience of the sales staff so that everyone has the same chance of getting a call.
- What is the expected years of experience of a person answering a new sales call?



EXAMPLE OF UNIFORM DISTRIBUTION

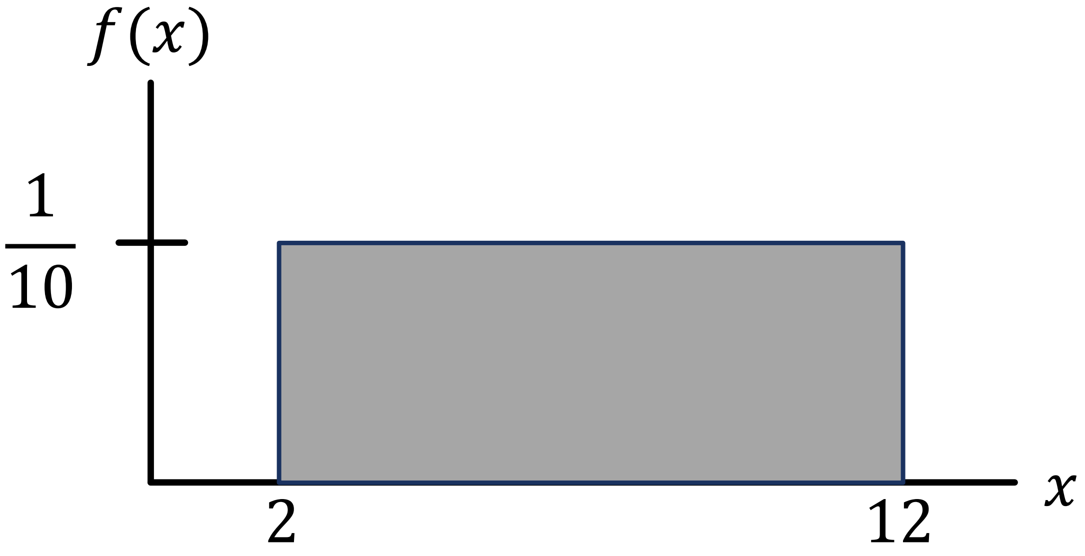
- Assume that sales calls that go into a company are uniformly distributed by the years of experience of the sales staff so that everyone has the same chance of getting a call.
- What is the expected years of experience of a person answering a new sales call?



$$E(x) = \frac{2+12}{2} = 7$$

EXAMPLE OF UNIFORM DISTRIBUTION

- Assume that sales calls that go into a company are uniformly distributed by the years of experience of the sales staff so that everyone has the same chance of getting a call.
- What is the expected years of experience of a person answering a new sales call?



$$E(x) = \frac{2+12}{2} = 7$$

$$Var(x) = \frac{(12-2)^2}{12} = 8.33$$

SUMMARY

- A random variable follows a uniform distribution whenever the probability is proportional to the interval's length.
- The probability density function for the uniform distribution is:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{elsewhere} \end{cases}$$

NORMAL DISTRIBUTION

DISTRIBUTIONS OF CONTINUOUS DATA

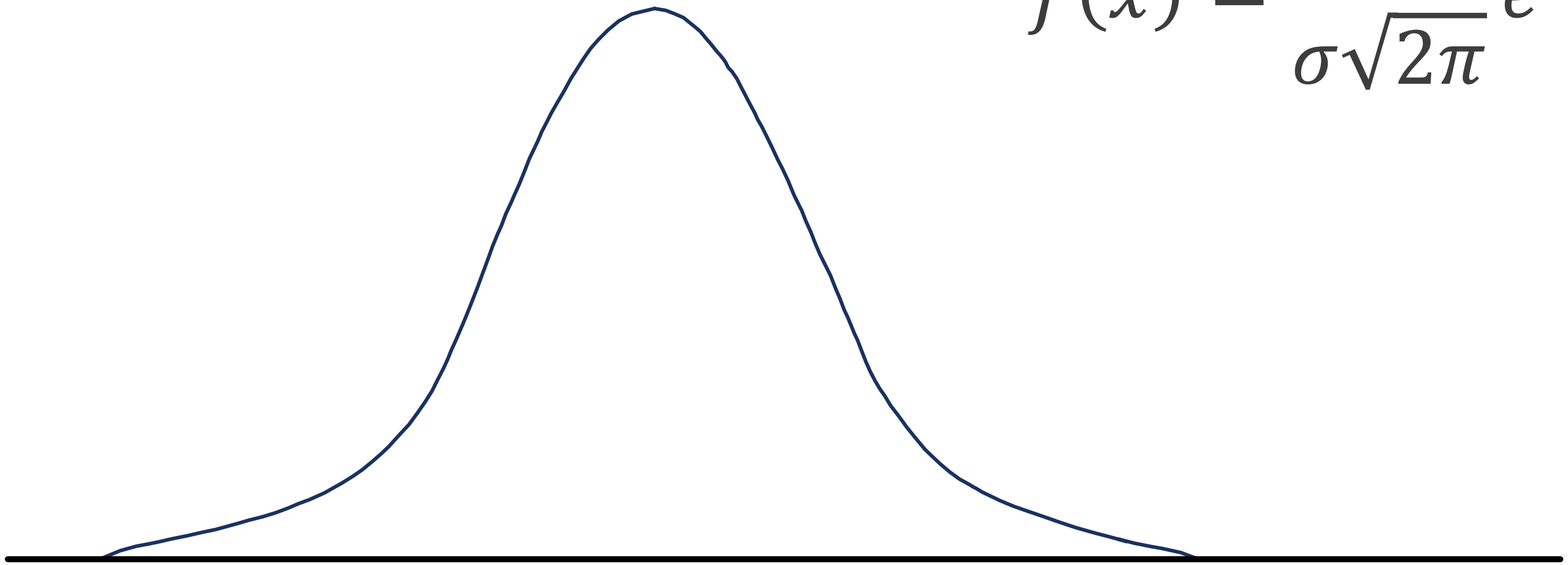
IMPORTANCE

- The **Normal probability distribution** is one of the most common and important distributions for describing a continuous random variable.
- The Normal distribution is the foundation of statistical inference:
  - Hypothesis Testing
  - Confidence Intervals
  - Regression Analysis
- Appears in nature and real-world data.

PROBABILITY DENSITY FUNCTION

- The probability density function for the Normal distribution is defined as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

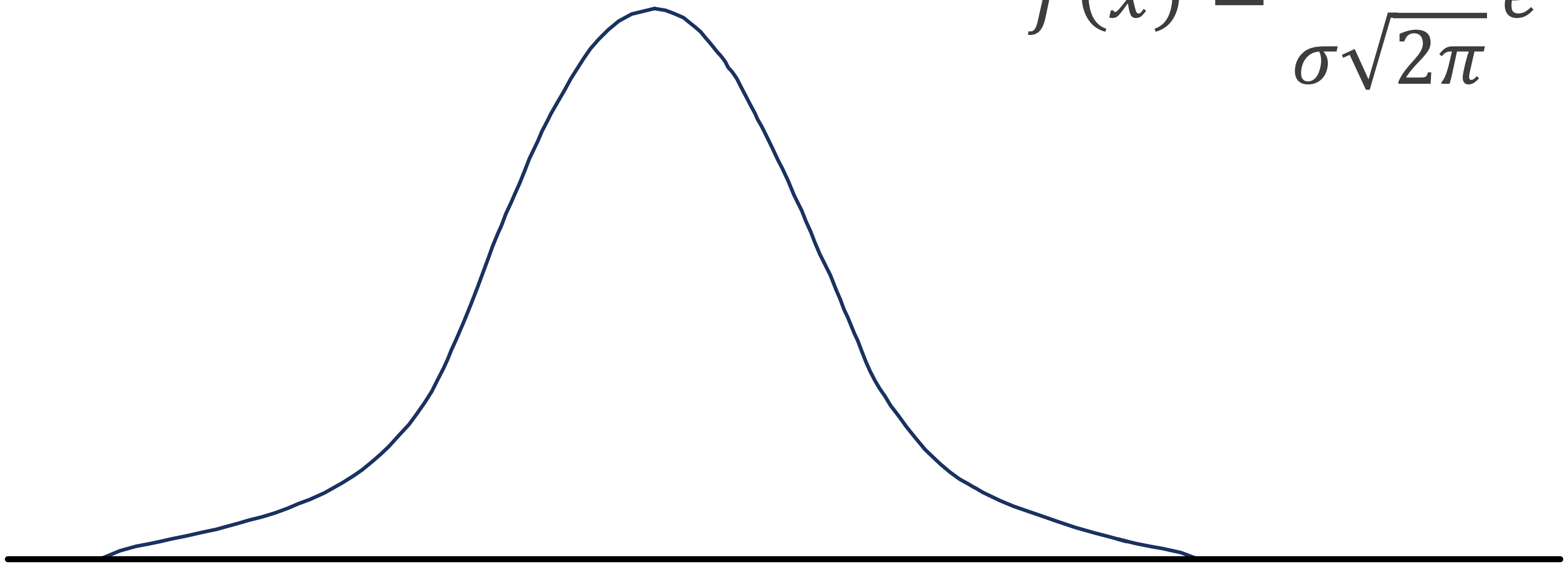


PROBABILITY DENSITY FUNCTION

- The probability density function for the Normal distribution is defined as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Mean = μ



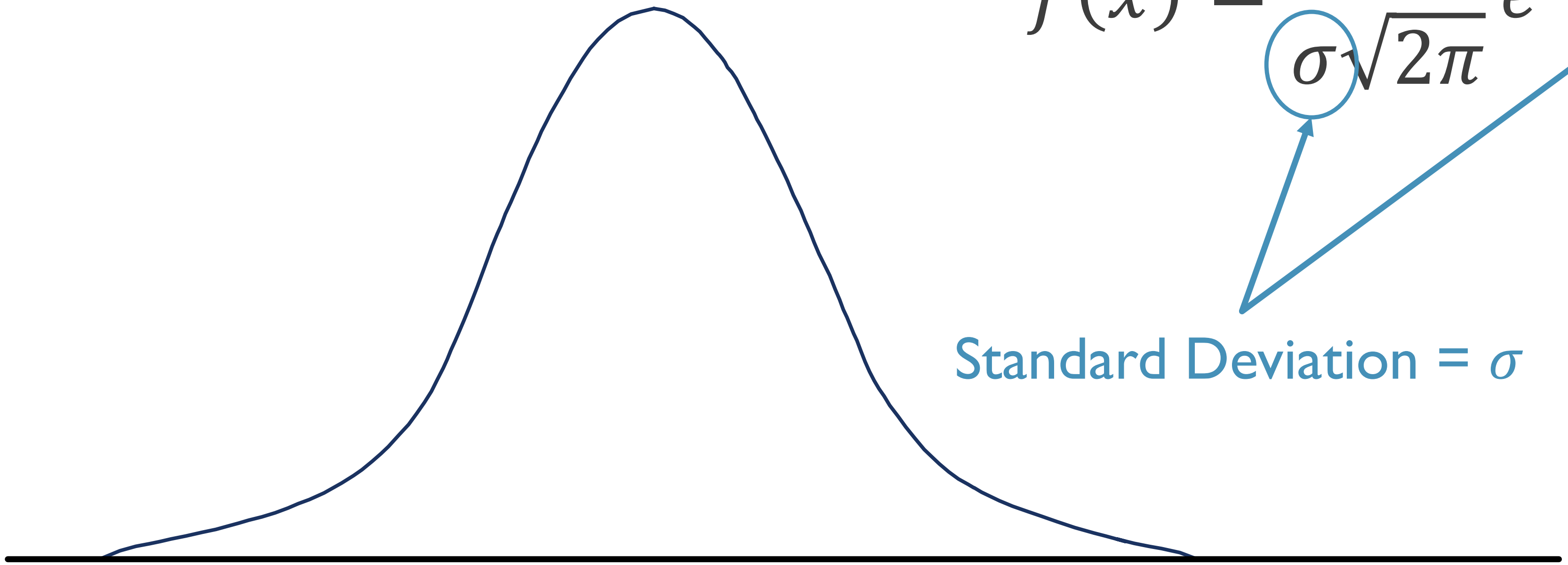
PROBABILITY DENSITY FUNCTION

- The probability density function for the Normal distribution is defined as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

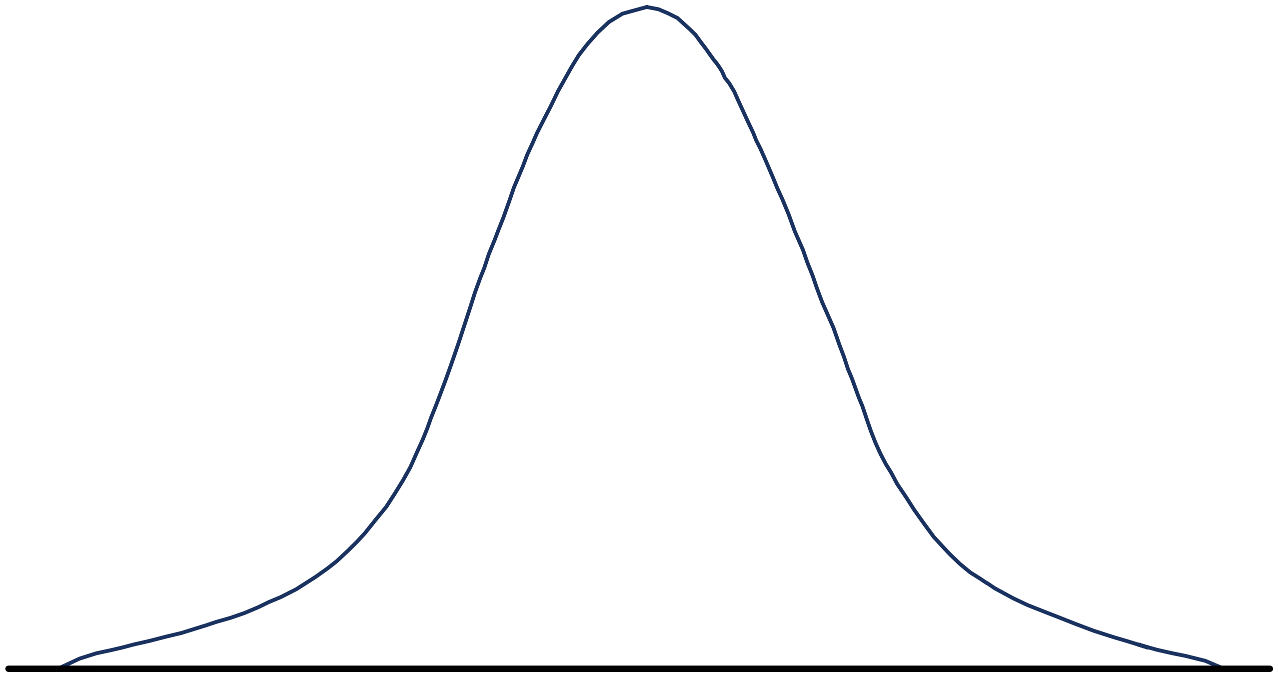
Mean = μ

Standard Deviation = σ



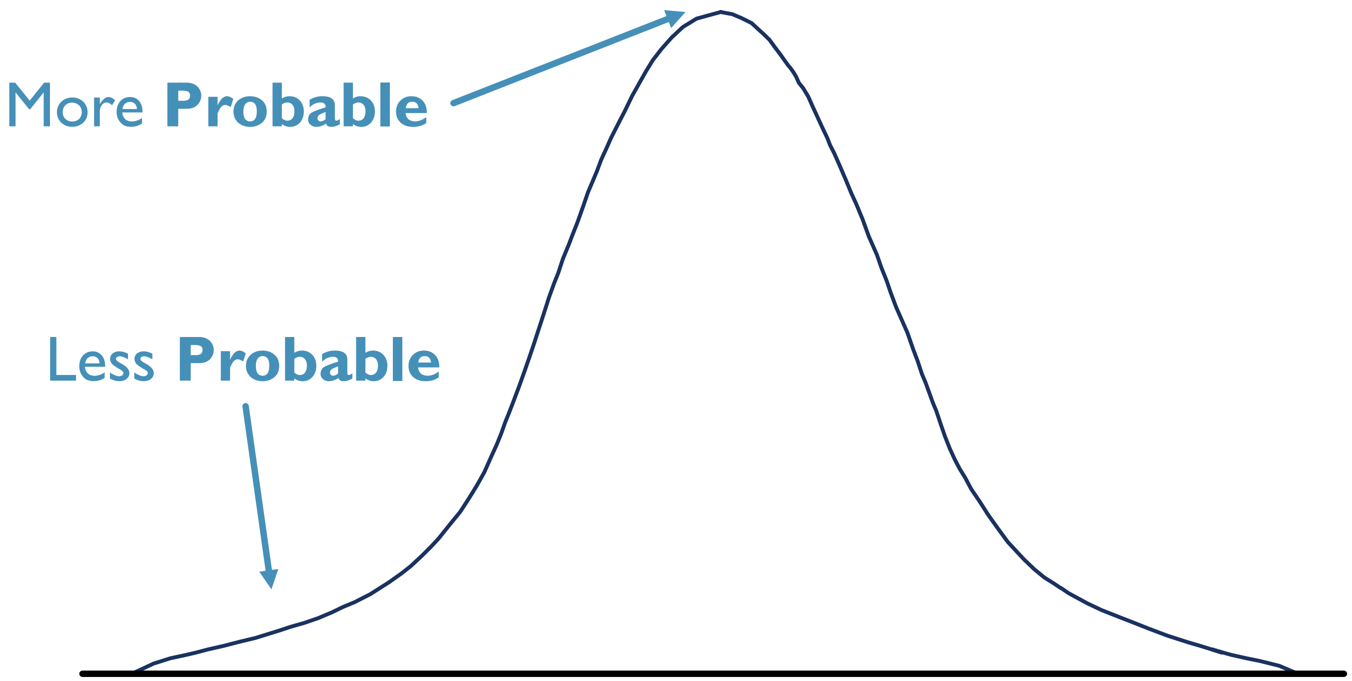
CHARACTERISTICS OF NORMAL DISTRIBUTION

- The Normal distribution has some useful characteristics:
  - Perfectly Symmetric (Skewness = 0)
  - Unimodal
  - Mean = Median = Mode
  - Asymptotic to x-axis (Can take any value from $-\infty$ to ∞)
  - Completely Defined by Mean and Standard Deviation



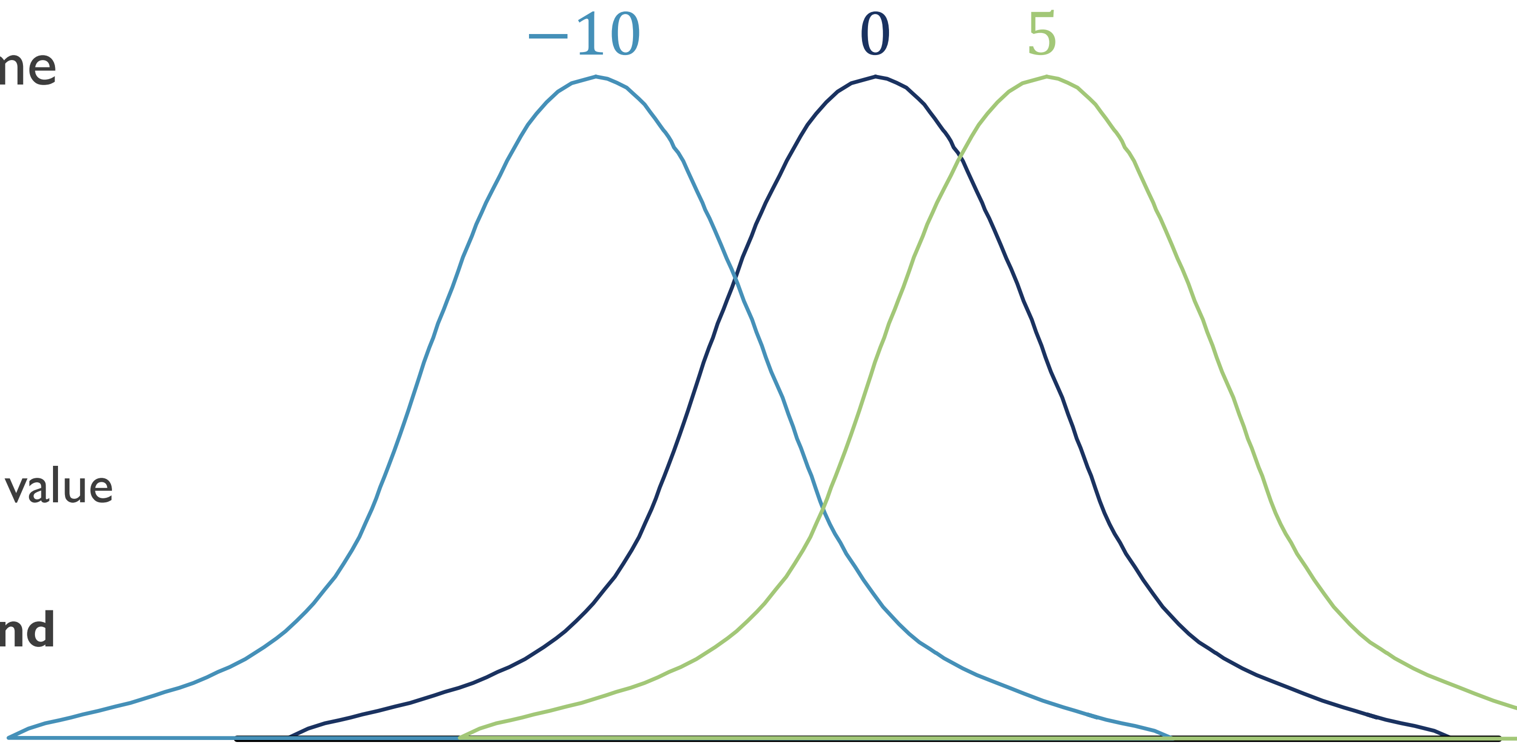
CHARACTERISTICS OF NORMAL DISTRIBUTION

- The Normal distribution has some useful characteristics:
  - Perfectly Symmetric (Skewness = 0)
  - Unimodal
  - Mean = Median = Mode
  - **Asymptotic to x-axis (Can take any value from $-\infty$ to ∞)**
  - Completely Defined by Mean and Standard Deviation



CHARACTERISTICS OF NORMAL DISTRIBUTION

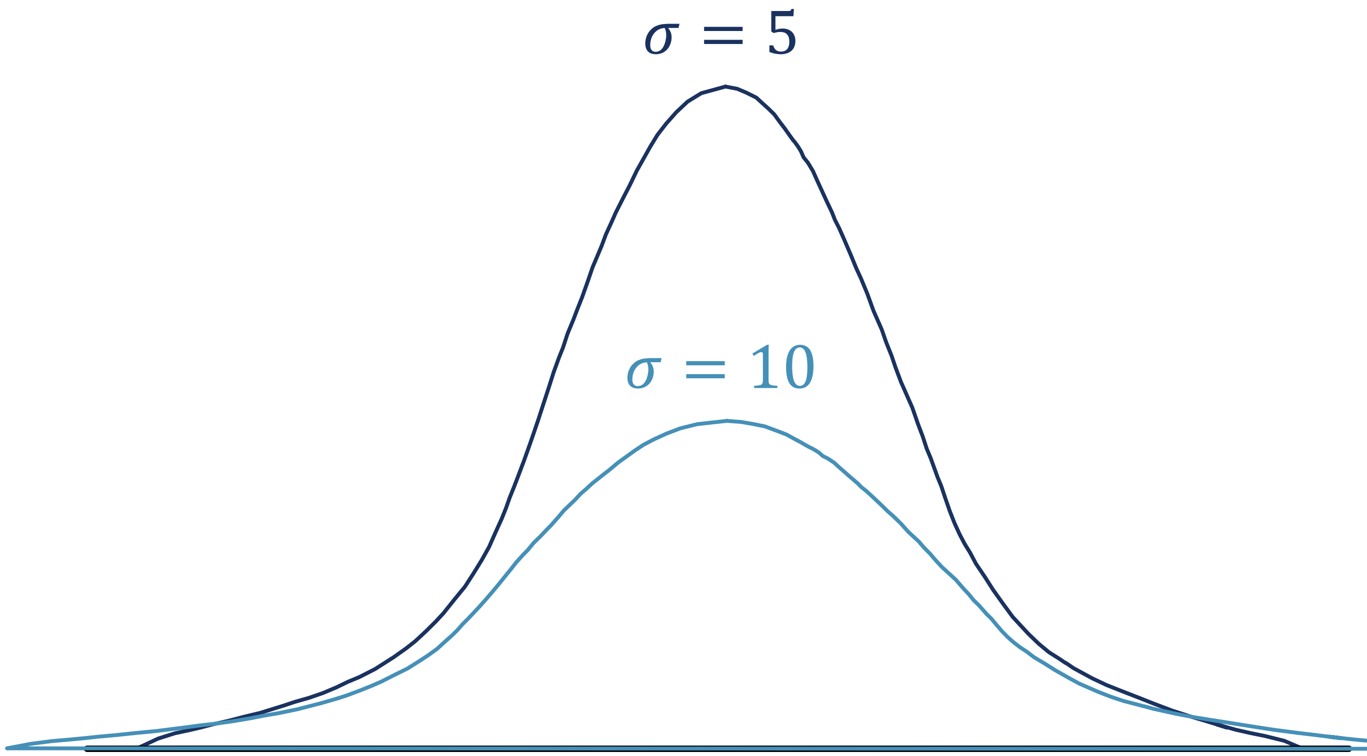
- The Normal distribution has some useful characteristics:
  - Perfectly Symmetric (Skewness = 0)
  - Unimodal
  - Mean = Median = Mode
  - Asymptotic to x-axis (Can take any value from $-\infty$ to ∞)
  - **Completely Defined by Mean and Standard Deviation**



Mean can take ANY value

CHARACTERISTICS OF NORMAL DISTRIBUTION

- The Normal distribution has some useful characteristics:
  - Perfectly Symmetric (Skewness = 0)
  - Unimodal
  - Mean = Median = Mode
  - Asymptotic to x-axis (Can take any value from $-\infty$ to ∞)
  - **Completely Defined by Mean and Standard Deviation**



Standard Deviation controls the width

SUMMARY

- The Normal probability distribution is one of the most common and important distributions for describing a continuous random variable.
- The Normal distribution is the foundation of statistical inference.
- The Normal distribution has some useful characteristics.

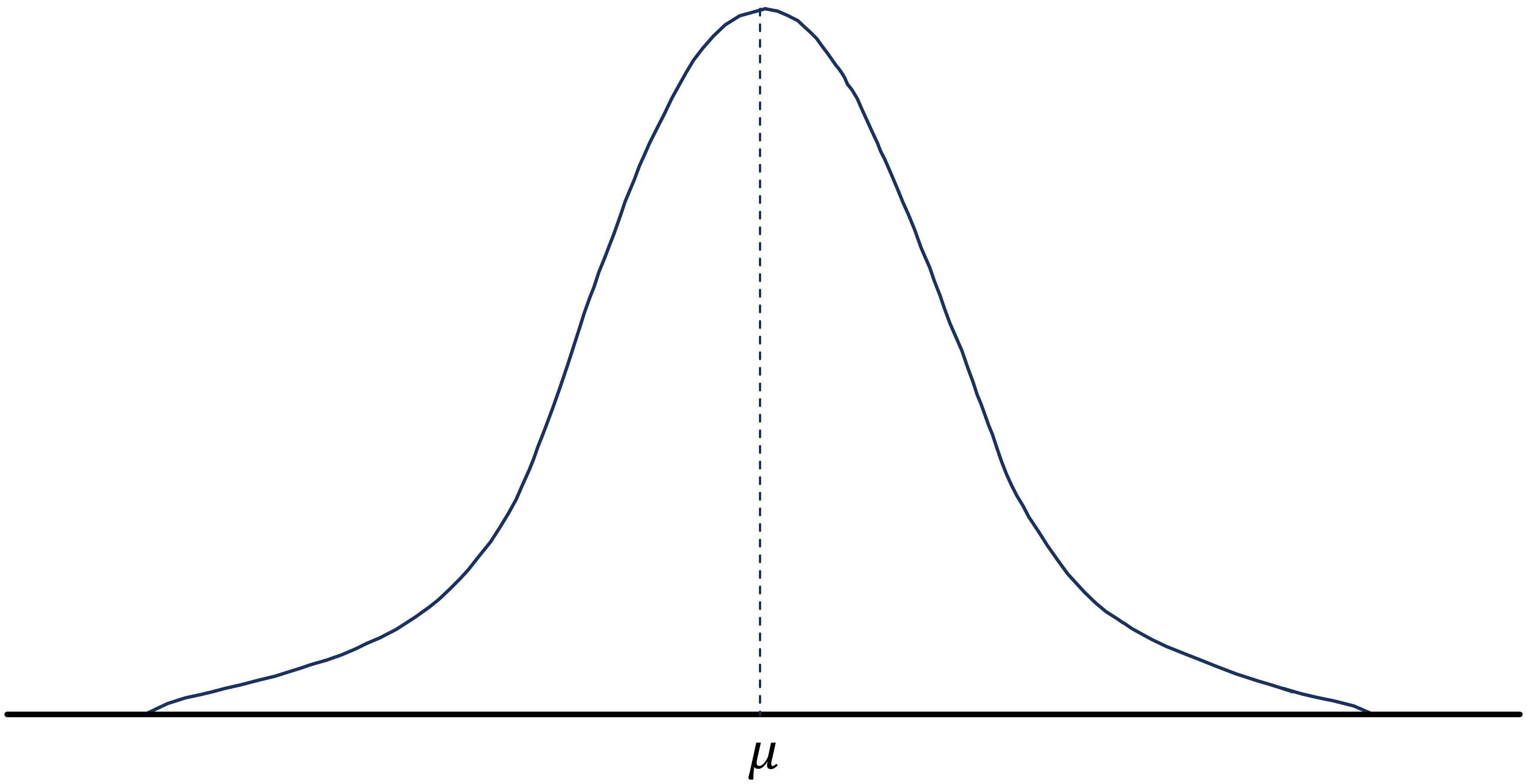
EMPIRICAL RULE

DISTRIBUTIONS OF CONTINUOUS DATA

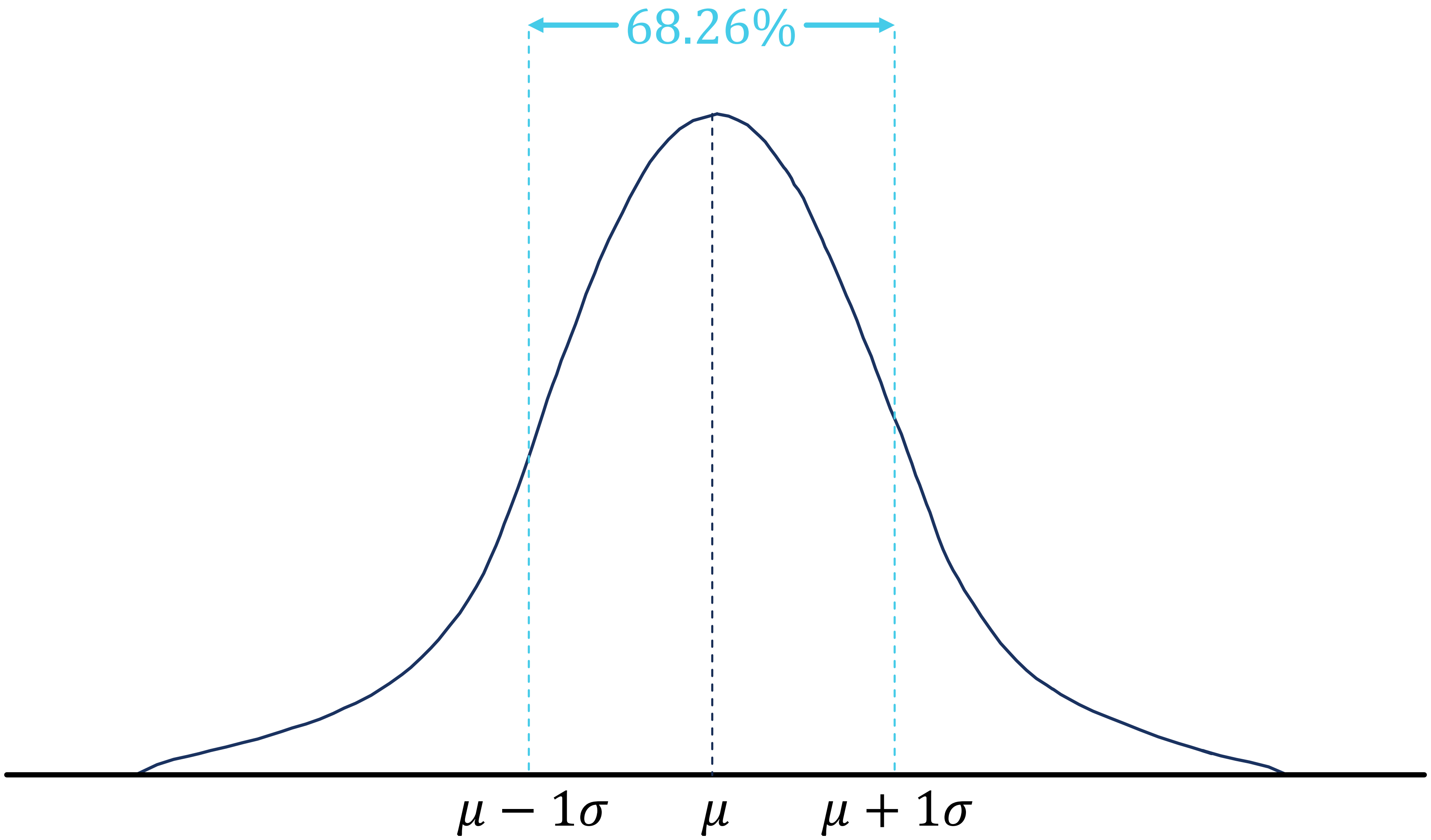
PROBABILITIES

- The probabilities for the Normal random variable are determined by the area under the curve.
- The **total area under the curve = 1**.
- Since the Normal distribution is perfectly symmetric around the mean (and median), then the area of the curve below the mean = above the mean = 0.5.

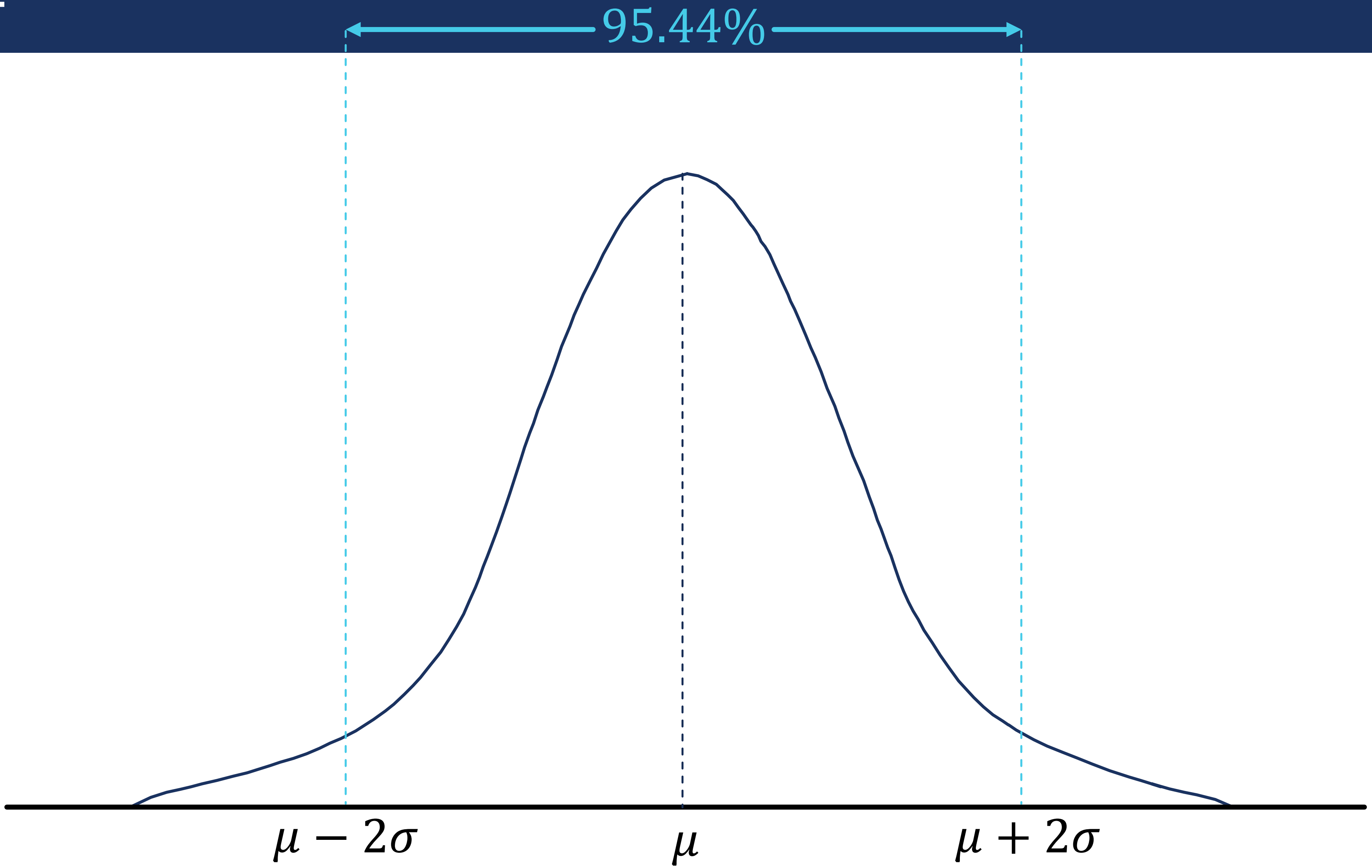
EMPIRICAL RULE



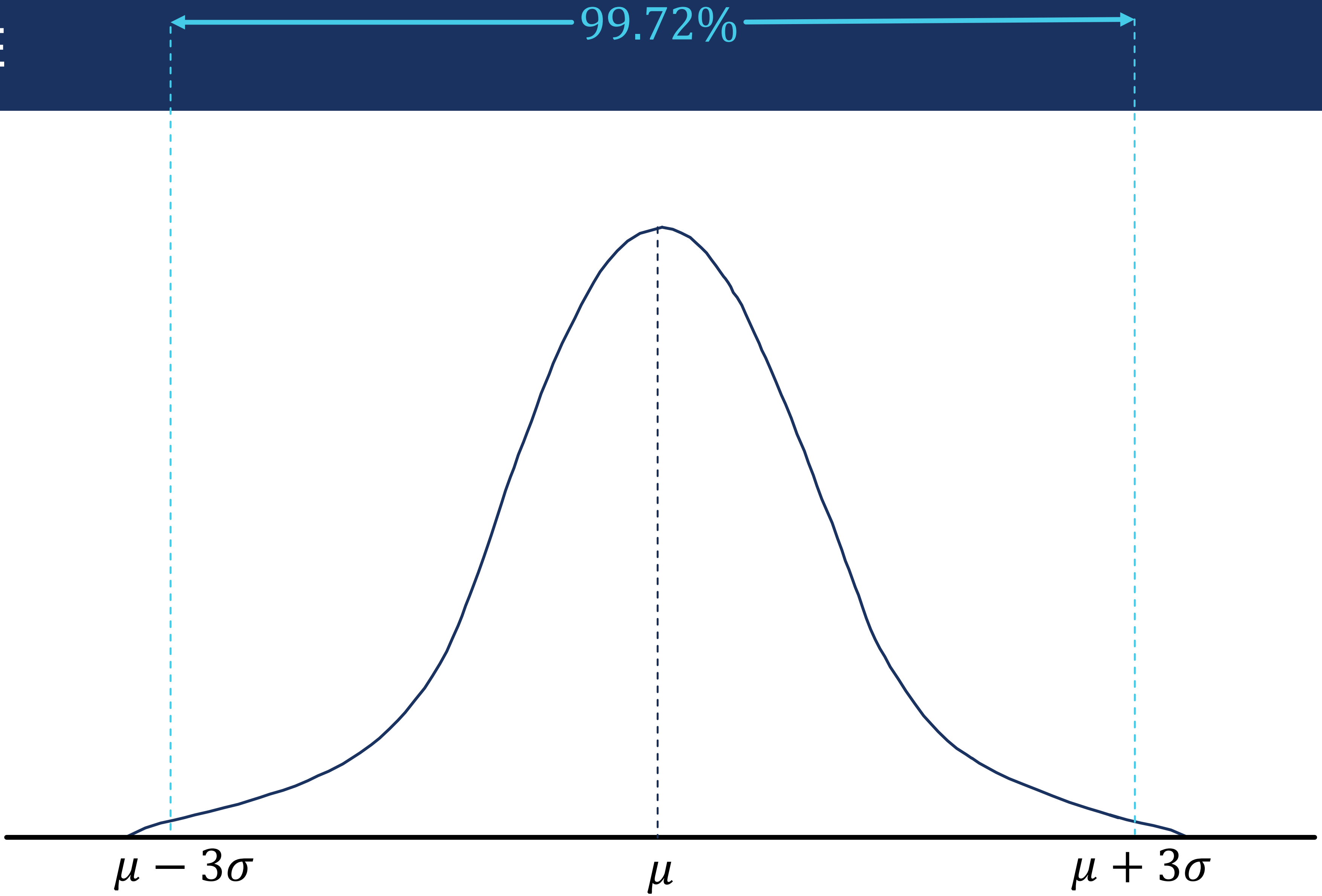
EMPIRICAL RULE



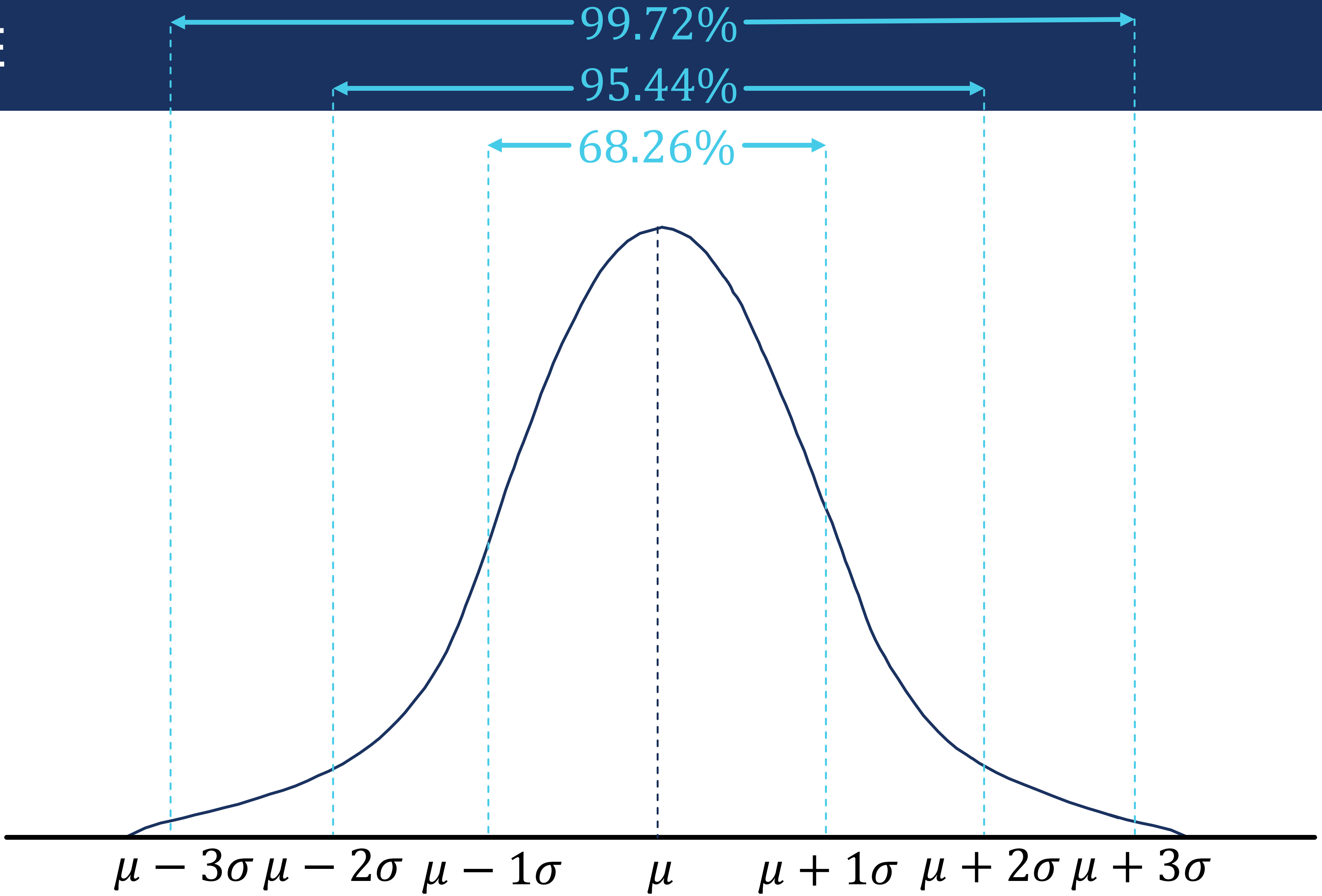
EMPIRICAL RULE



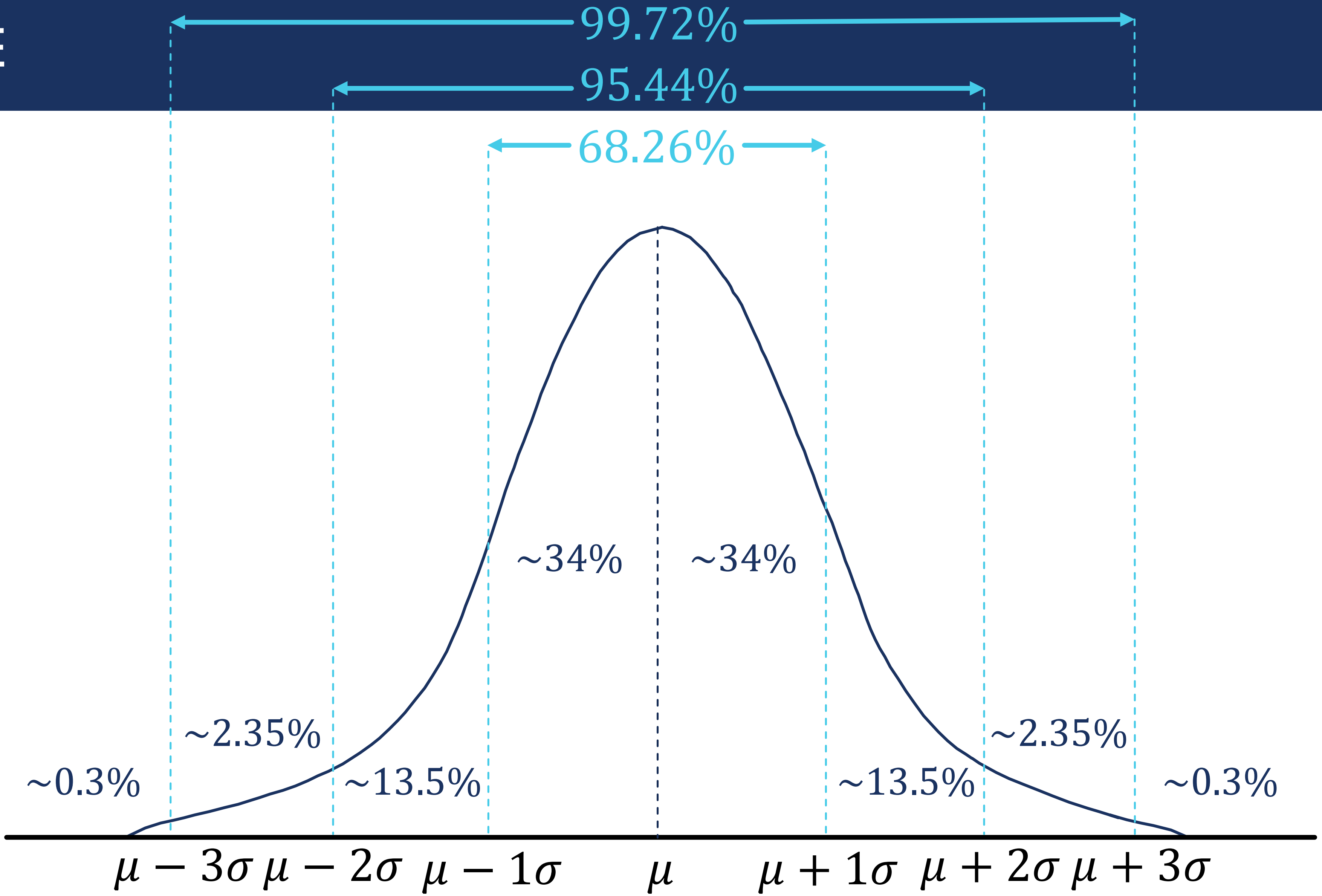
EMPIRICAL RULE



EMPIRICAL RULE



EMPIRICAL RULE

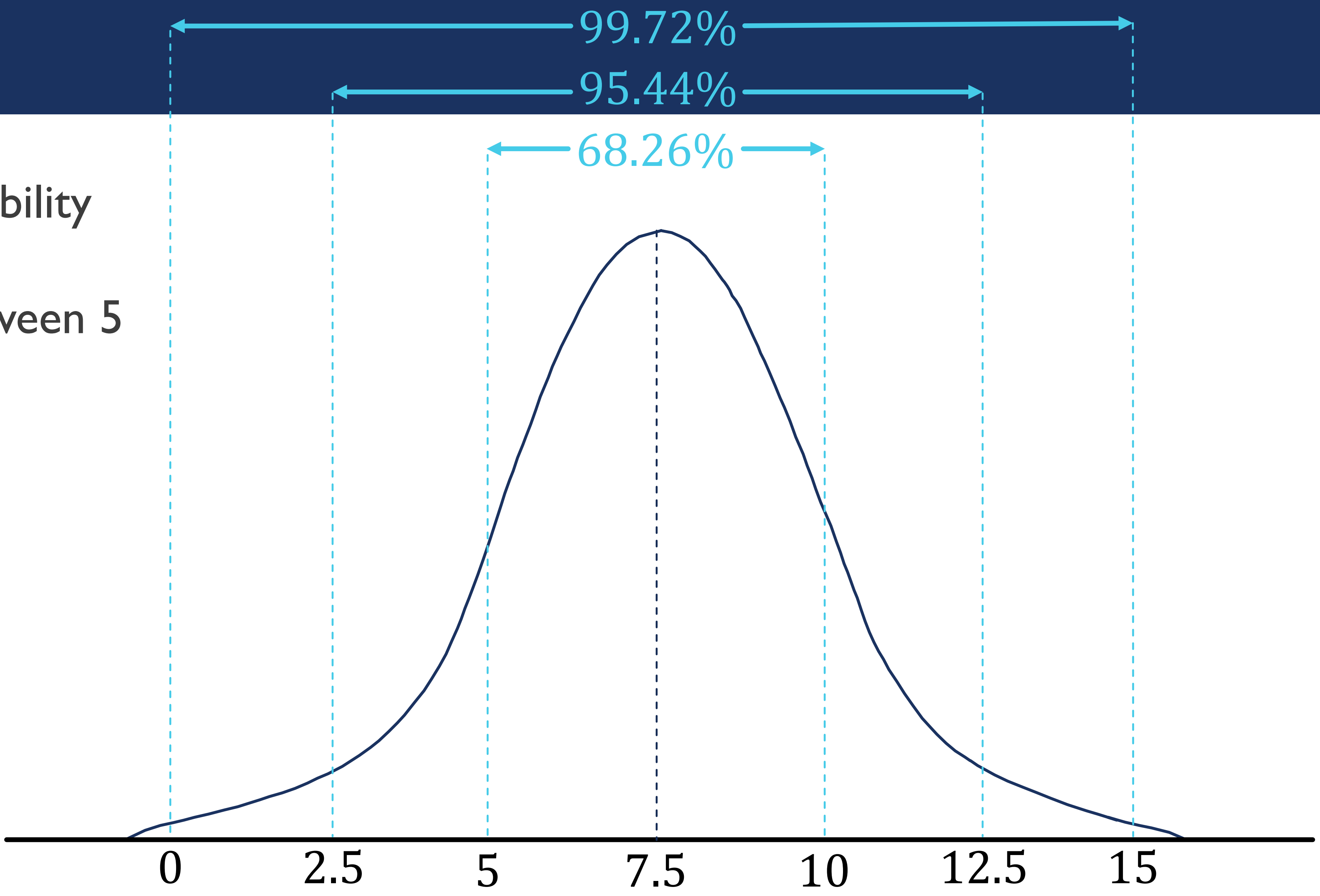


EXAMPLE

- Assume new employees at a company have previous years of professional experience that follow a Normal distribution where the mean is 7.5 and the standard deviation is 2.5.
- What is the probability any random new employee has between 5 and 10 years of experience?

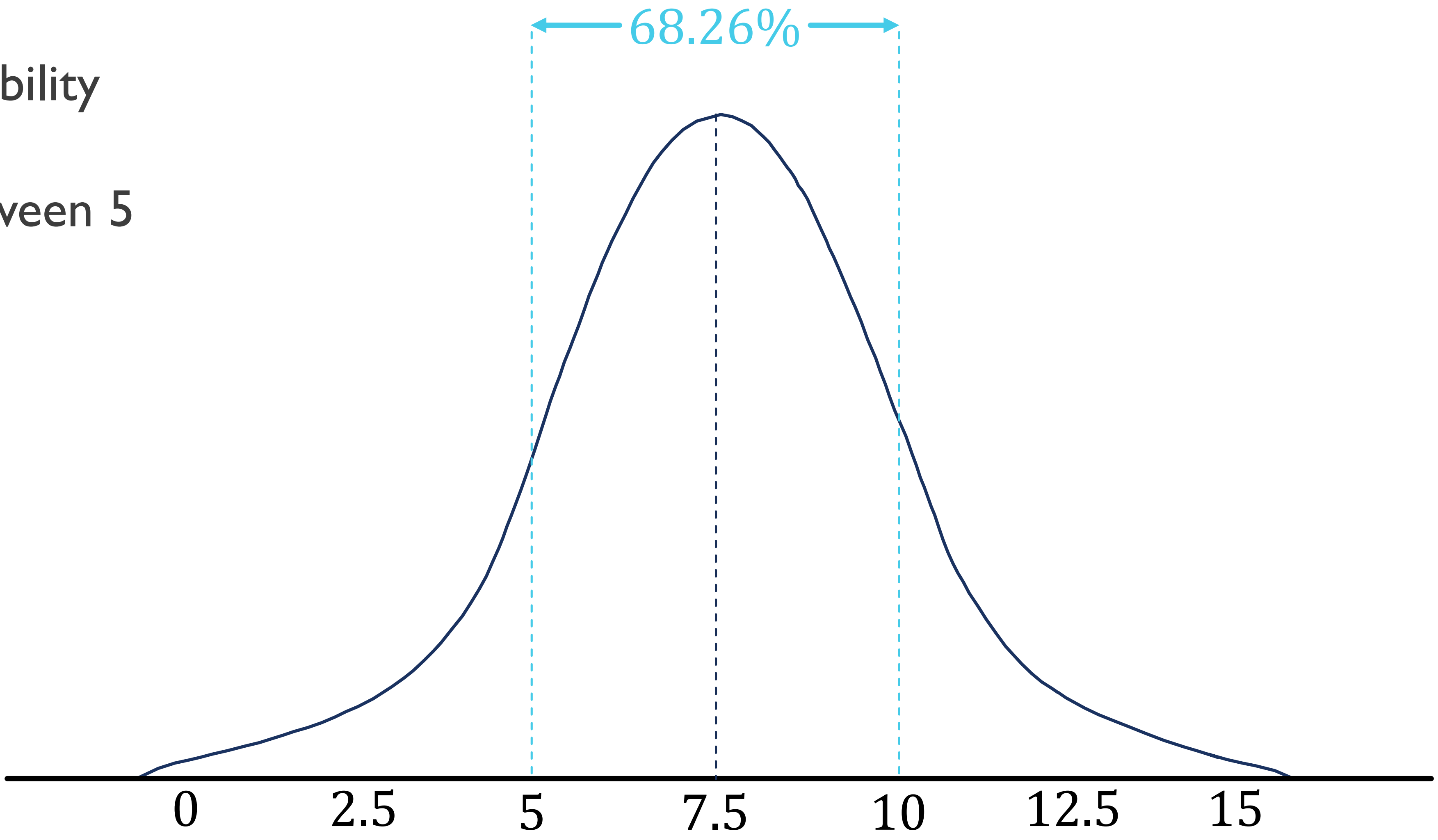
EXAMPLE

- What is the probability any random new employee has between 5 and 10 years of experience?



EXAMPLE

- What is the probability any random new employee has between 5 and 10 years of experience?

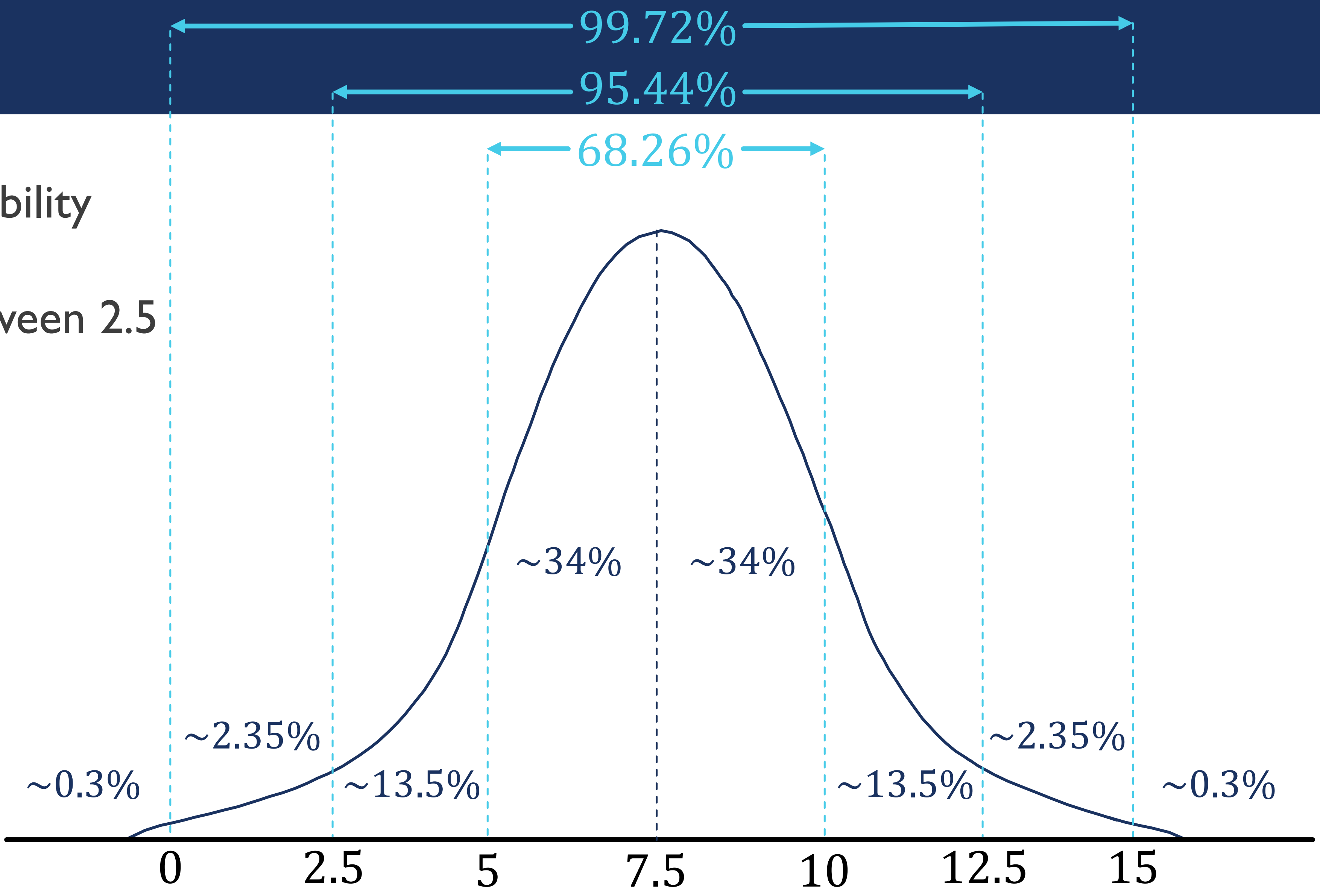


EXAMPLE

- Assume new employees at a company have previous years of professional experience that follow a Normal distribution where the mean is 7.5 and the standard deviation is 2.5.
- What is the probability any random new employee has between 2.5 and 10 years of experience?

EXAMPLE

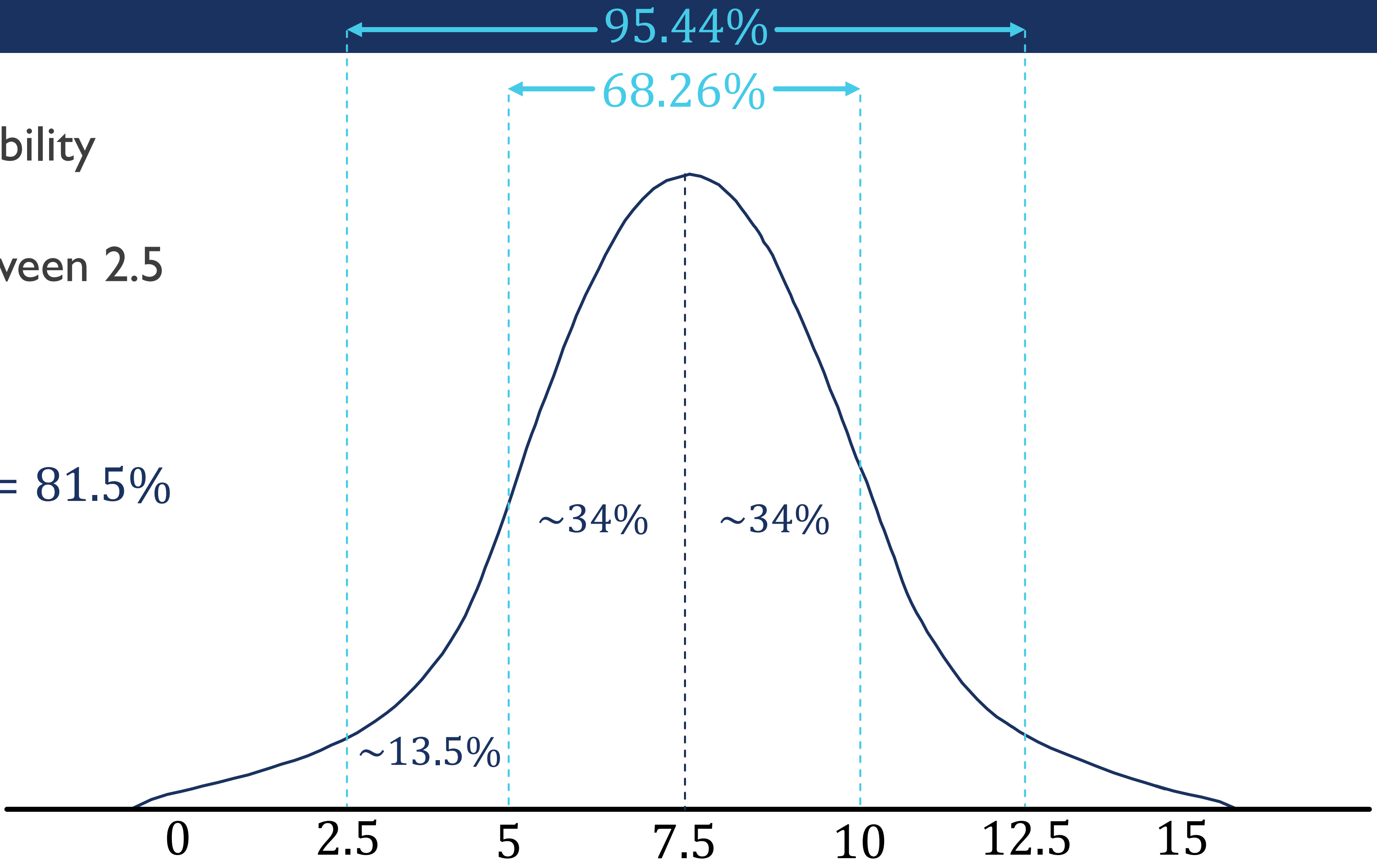
- What is the probability any random new employee has between 2.5 and 10 years of experience?



EXAMPLE

- What is the probability any random new employee has between 2.5 and 10 years of experience?

$$34 + 34 + 13.5 = 81.5\%$$



SUMMARY

- The empirical rule (68, 95, 99.7 rule) is good for quick, fast, rough analysis.
- Not good for exact analysis unless your interests are only in the integer standard deviations.
- What about fractions of standard deviations away from the mean?
- Need another way to quickly calculate area under the curve.

STANDARD SCORES

DISTRIBUTIONS OF CONTINUOUS DATA

CONVERSION OF NORMAL DISTRIBUTIONS

- A random variable having a Normal distribution with a mean of 0 and a standard deviation of 1 is said to have a **standard Normal probability distribution**.
- All Normal distributions can be converted into standard Normal distributions for ease of computing probabilities under the curve.
- Standard Normal probability tables help calculate area under the curve.

STANDARD NORMAL TABLE

- The standard Normal table is an extension of the empirical rule where the area under the standard Normal curve to the left of any point is calculated up to two decimal points.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7517	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
.	.	.	.	.	.	.	.	.	.	.

STANDARD NORMAL TABLE

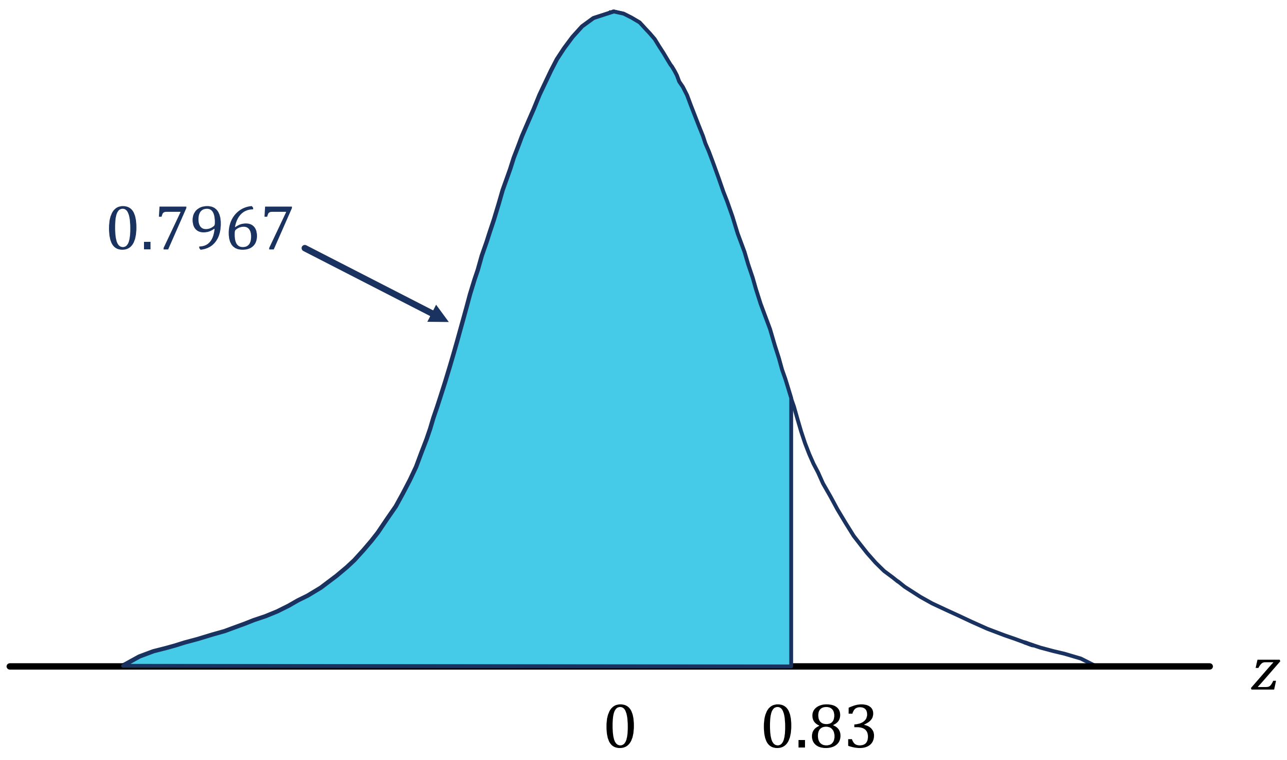
- The standard Normal table is an extension of the empirical rule where the area under the standard Normal curve to the left of any point is calculated up to two decimal points.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.		.	.	.	.	.	.
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7517	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
.	.	.	.	.	.	.	.	.	.	.

$P(z \leq 0.83)$

STANDARD NORMAL TABLE

- The standard Normal table is an extension of the empirical rule where the area under the standard Normal curve to the **left** of any point is calculated up to two decimal points.



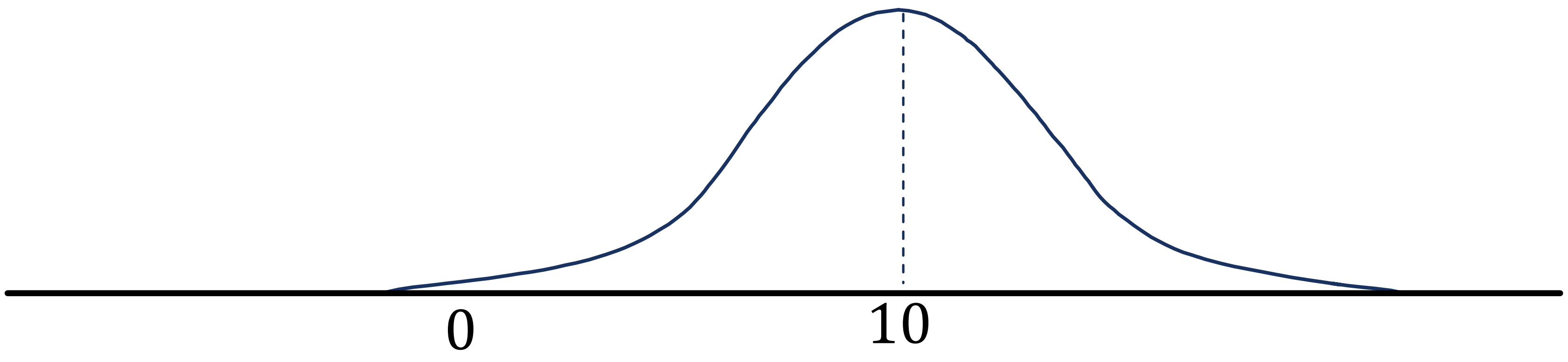
CALCULATING OPPOSITE PROBABILITIES

- The standard Normal table is an extension of the empirical rule where the area under the standard Normal curve to the **left** of any point is calculated up to two decimal points.
- To calculate values to the **right** of any point, use the laws of probability:

$$\begin{aligned} P(z > 0.83) &= 1 - P(z \leq 0.83) \\ &= 1 - 0.7967 \\ &= 0.2033 \end{aligned}$$

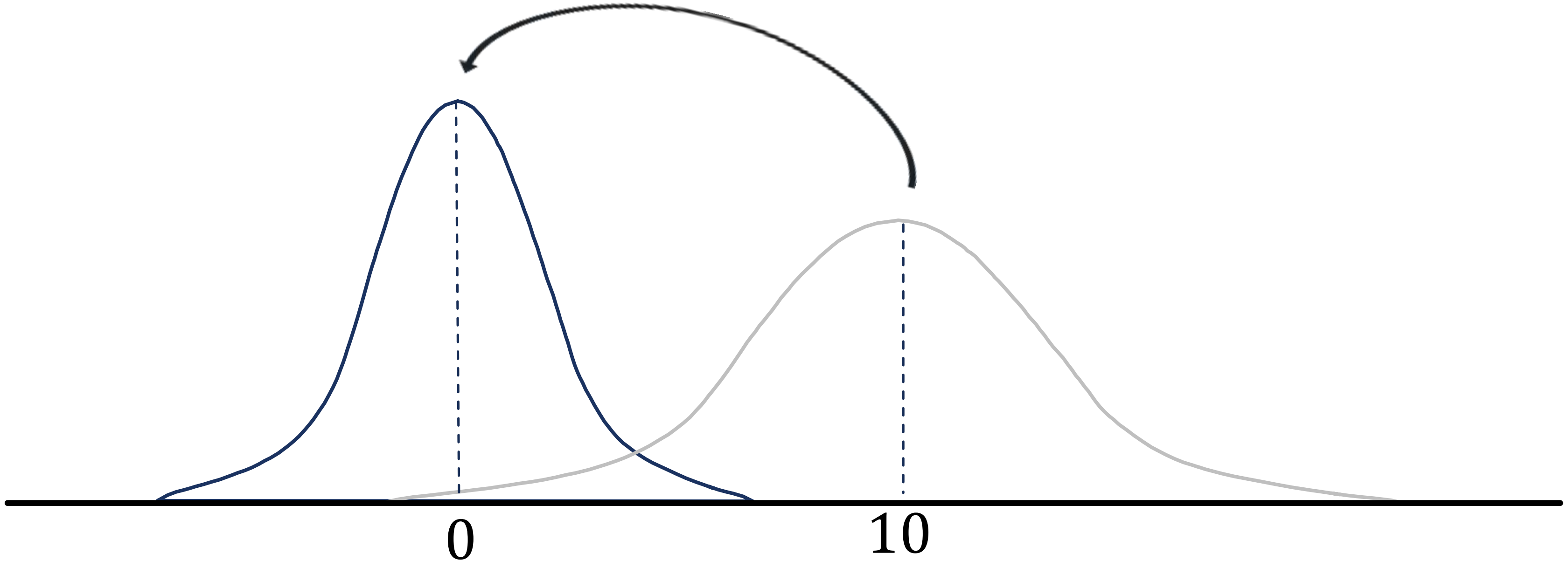
CONVERSION OF NORMAL DISTRIBUTIONS

- All Normal distributions can be converted into standard Normal distributions for ease of computing probabilities under the curve.



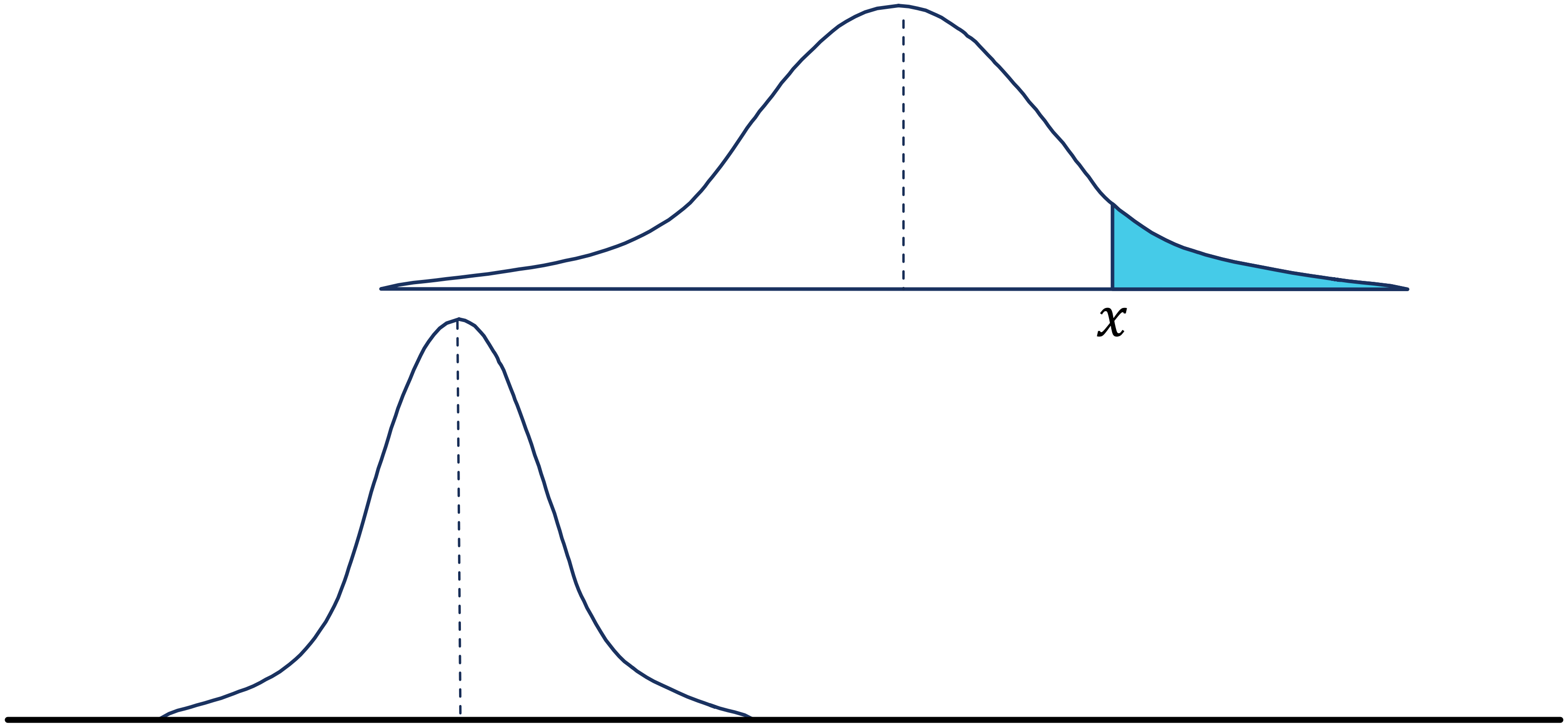
CONVERSION OF NORMAL DISTRIBUTIONS

- All Normal distributions can be converted into standard Normal distributions for ease of computing probabilities under the curve.



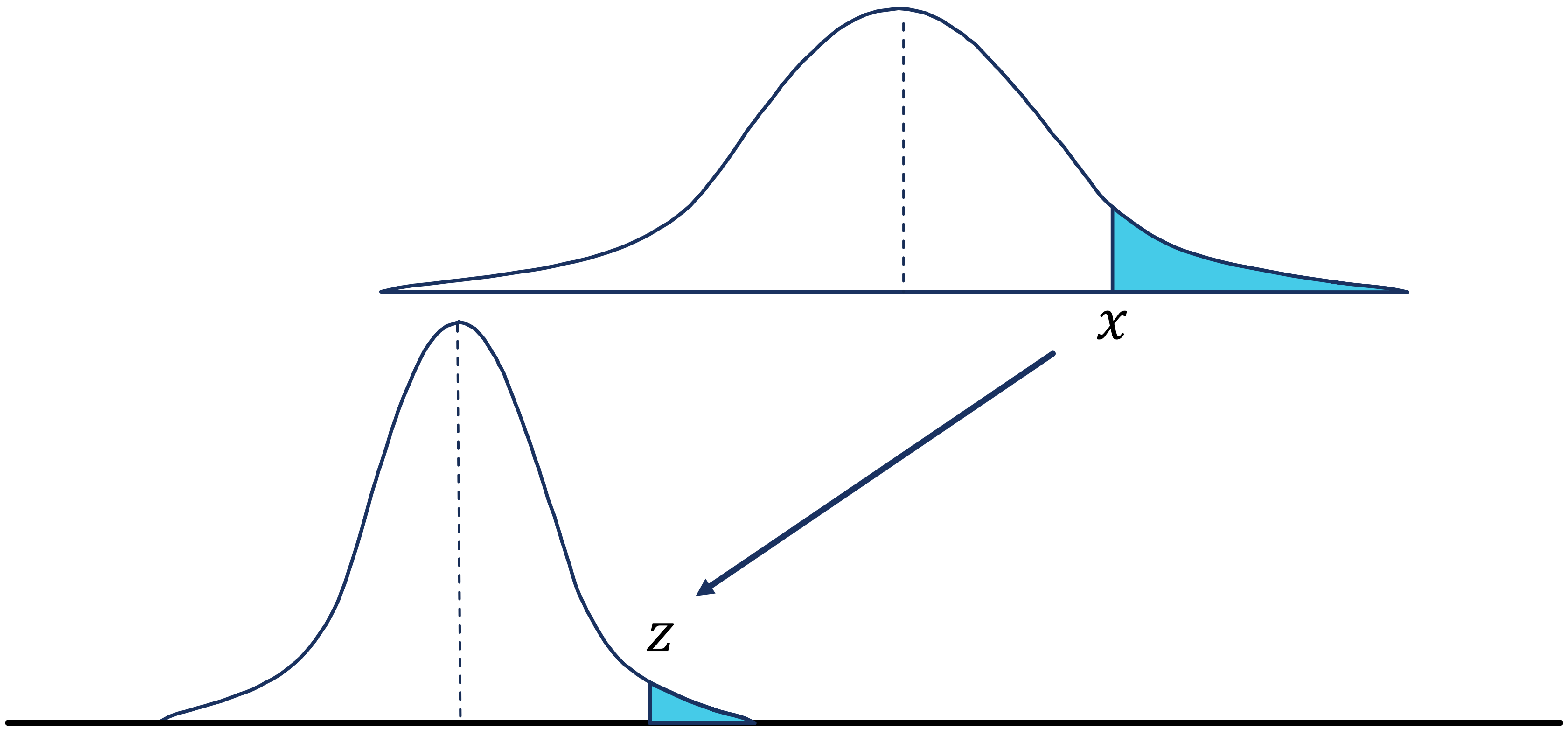
CONVERSION OF NORMAL DISTRIBUTIONS

- All Normal distributions can be converted into standard Normal distributions for ease of computing probabilities under the curve.



CONVERSION OF NORMAL DISTRIBUTIONS

- All Normal distributions can be converted into standard Normal distributions for ease of computing probabilities under the curve.



Z-SCORES

- Converting from any point on any Normal distribution (with mean μ and standard deviation σ) to the corresponding point on the standard Normal distribution can be done through **z-scores**.

$$z = \frac{x - \mu}{\sigma}$$

Z-SCORES BIKE DATA EXAMPLE

- Assume that the daily number of total users follows a Normal distribution. The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that any random day has more than 6,000 total users?

Z-SCORES BIKE DATA EXAMPLE

- Assume that the daily number of total users follows a Normal distribution. The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that any random day has more than 6,000 total users?

$$z = \frac{x - \mu}{\sigma} = \frac{6{,}000 - 4{,}504}{1937} = 0.77$$

Z-SCORES BIKE DATA EXAMPLE

$$z = \frac{x - \mu}{\sigma} = \frac{6{,}000 - 4{,}504}{1937} = 0.77$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7517	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
.	.	.	.	.	.	.	.	.	.	.

$P(z \leq 0.77)$

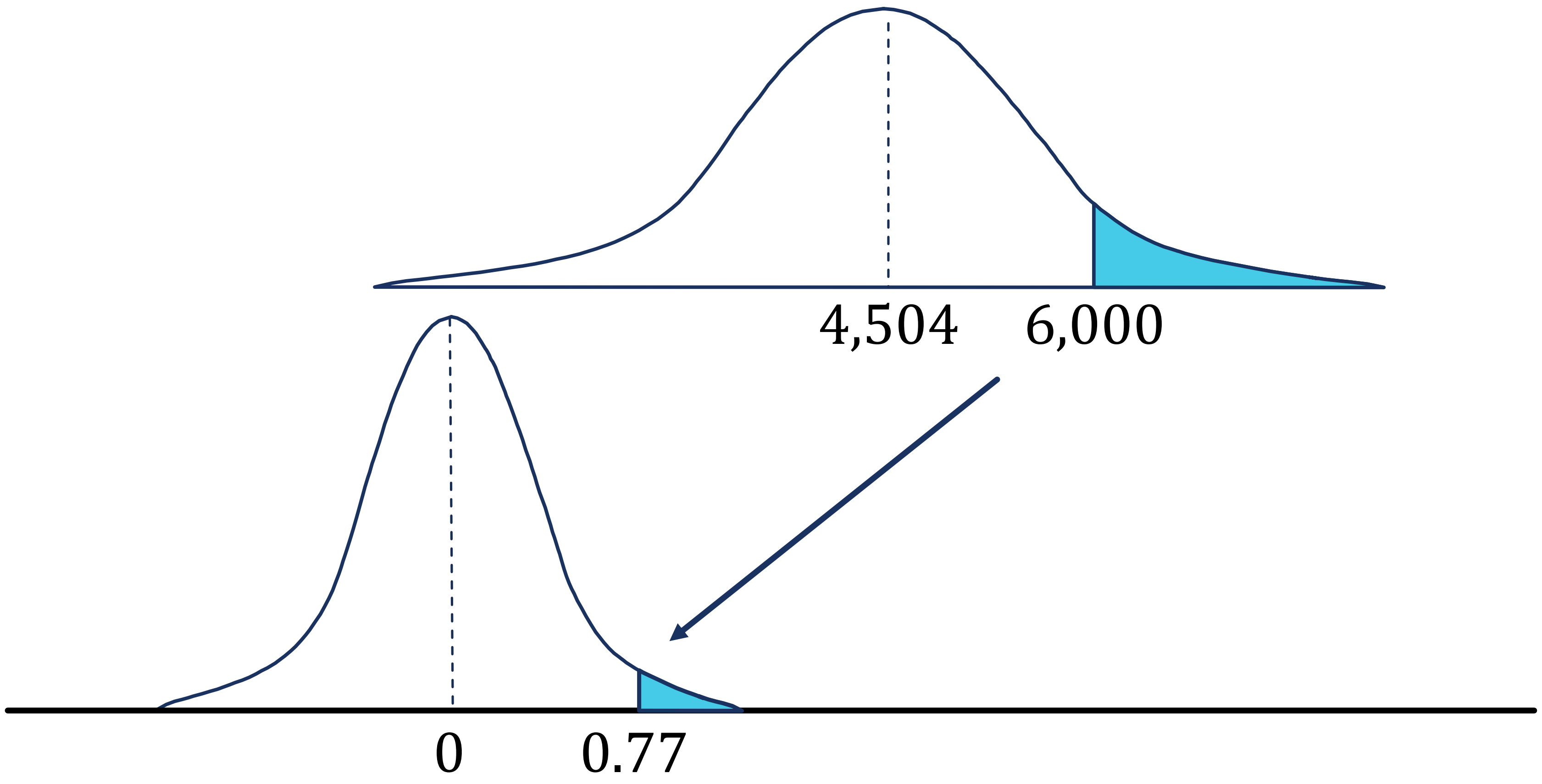
Z-SCORES BIKE DATA EXAMPLE

- Assume that the daily number of total users follows a Normal distribution. The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that any random day has more than 6,000 total users?

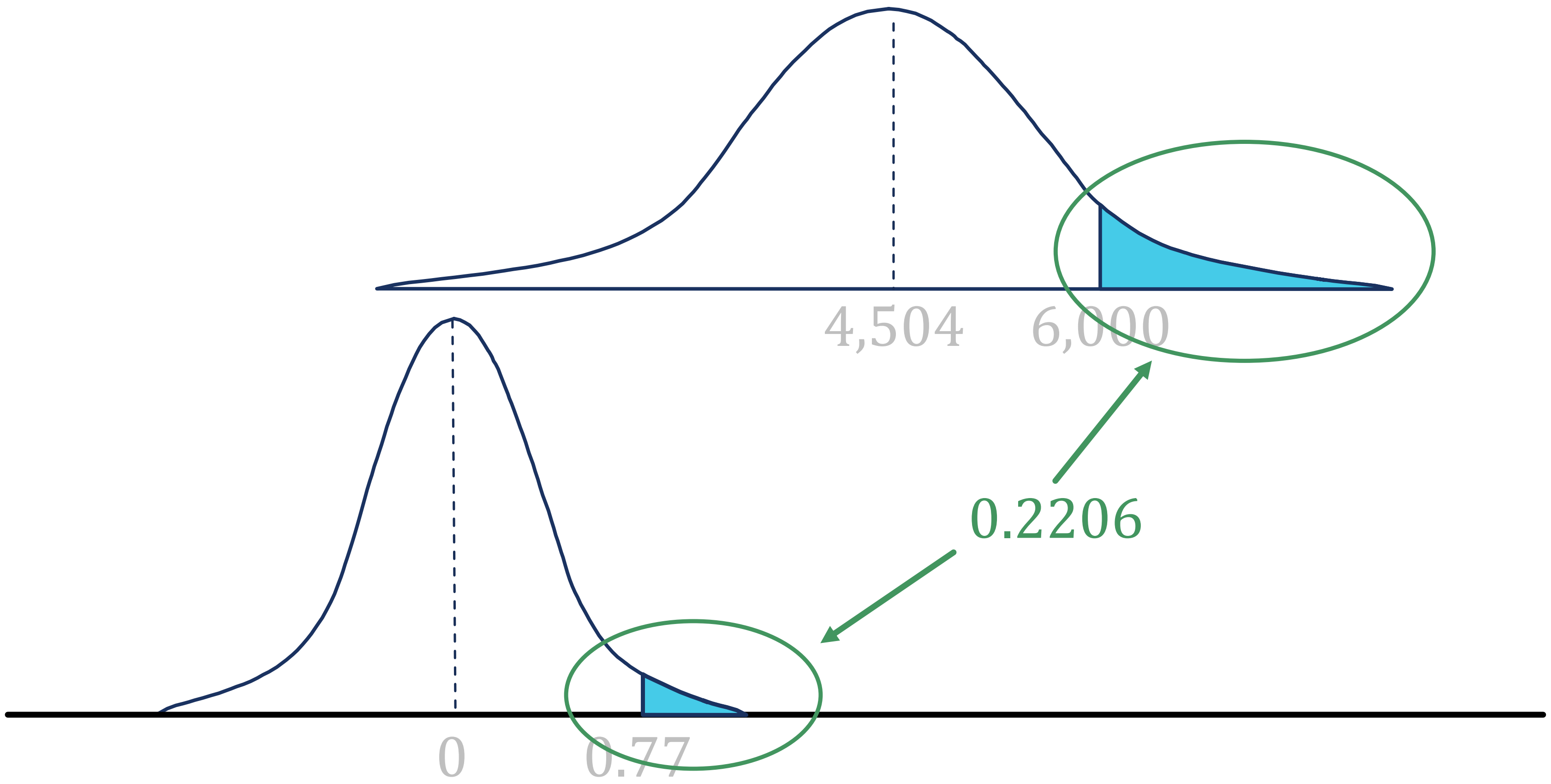
$$z = \frac{x - \mu}{\sigma} = \frac{6{,}000 - 4{,}504}{1937} = 0.77$$

$$\begin{aligned} P(z > 0.77) &= 1 - P(z \leq 0.77) \\ &= 1 - 0.7794 \\ &= 0.2206 \end{aligned}$$

Z-SCORES BIKE DATA EXAMPLE



Z-SCORES BIKE DATA EXAMPLE



Z-SCORES BIKE DATA EXAMPLE

- Assume that the daily number of total users follows a Normal distribution. The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the number of daily users that would be in the bottom 10% of daily users?

Z-SCORES BIKE DATA EXAMPLE

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
.	.	.	.	.	.	.	.	.	.	.

$$P(z \leq ?) = 0.1$$

Z-SCORES BIKE DATA EXAMPLE

- Assume that the daily number of total users follows a Normal distribution. The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the number of daily users that would be in the bottom 10% of daily users?

$$z = \frac{x - \mu}{\sigma} = -1.28 = \frac{x - 4{,}504}{1{,}937}$$

$$x = 2{,}024.64$$

SUMMARY

- A random variable having a Normal distribution with a mean of 0 and a standard deviation of 1 is said to have a **standard Normal probability distribution**.
- All Normal distributions can be converted into standard Normal distributions for ease of computing probabilities under the curve.
- Standard Normal probability tables help calculate area under the curve.